Beszéd-zene lejátszási listák nyelvtechnológiai vonatkozása

Benyeda Ivett[1], Jani Mátyás[2], Lukács Gergely[2]

[1] MTA Nyelvtudományi Intézet,
1068, Budapest, Benczúr utca 33.
benyeda.ivett@nytud.mta.hu
[2] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
{jani.matyas, lukacs}@itk.ppke.hu

Kivonat: Az internetes és okostelefonos médiafogyasztás lehetővé és szükségessé teszi a tartalom személyre szabását. Hangalapú média esetén ezzel a lejátszási lista (playlist generation) témakör foglalkozik. A korábbi munkák a területen kizárólag a zene alapú lejátszási listákkal foglalkoztak, a beszéd-zene lejátszási listákkal foglalkozó első kutatások is az akusztikai oldalt vizsgálták. Jelen munka, úttörő módon, a beszéd-zene lejátszási listák készítésének nyelvtechnológiai oldalával foglalkozik. Az előzetes vizsgálatok alapján javaslatot tesz a beszéd-zene lejátszási lista készítésének vázára. A nyelvtechnológiai feldolgozásnál különösen a hangulati, érzelmi vonatkozásnak, ezek dalszövegekből, interjúátiratokból és hangzó beszédből való hangulatkinyerésének van jelentősége. Ehhez hangulati szótárakat használunk fel, hangulati szavak dalszövegekben és interjúátiratokban való előfordulását vizsgáljuk. A beszédet tartalmazó hanganyagok esetén a szöveg előállításához automatikus beszédfelismerést is végzünk, kétféle módon: a teljes hanganyag felismerésével, ill. a hangulati szavakra való fókuszálással. Vizsgáljuk, hogy a hangulati szavak előfordulását hogyan változtatja meg a beszédfelismerés korlátozott minősége. A munkát angol nyelvű szótárakkal és BBC anyagokon végeztük.

1 Bevezetés

Az Internet alapvetően megváltoztatta a médiafogyasztási lehetőségeket és szokásokat, mivel a felhasználók számára elérhető médiatartalom robbanásszerűen növekedett. Az egyéni válogatás időigénye és nehézkessége miatt szükségessé vált a médiatartalmat személyre szabni. A személyre szabott tartalom függhet az egyéni érdeklődéstől, az aktuális szituációtól, de a felhasználó (médiafogyasztó) közösségi kapcsolatoktól is. Az okostelefonok és a mobilinternet jelenlegi terjedése ezeket a trendeket – médiatartalmak elérhetőségét, az egyszerű kezelés iránti igényt és a személyre szabás lehetőségét – tovább erősíti.

A hangalapú médiát több szempont miatt is kiemelt jelentőségűnek látjuk: jóval nagyobb az információtartalma az írásos anyagokénál, a befogadáshoz szükséges erőfeszítés ugyanakkor jóval kisebb [1] és megengedi az egyidejű fizikai aktivitást, pl. a közlekedést, a házi vagy a házkörüli fizikai munkát. Ezt látszik alátámasztani a

személyre szabott online zenei rádiók, pl. Pandora, Spotify vagy last.fm megjelenése és gyors növekedése. A szokásos közösségi médiához képest érdekes alternatívát jelent a hangalapú közösségi média is. Az előzménynek tekinthető kisközösségi rádiók (angol: small community radio, low power radio station) közösségépítő hatására számos példa van a világban, az okostelefonos hangalapú közösségi média prototípusa egy kontrollált kísérletben 30-60%-kal növelte az emberi kapcsolatok számát [2].

A tartalom személyre szabása a hang- (vagy videó-) alapú folyamatos tartalomszolgáltatásnál nagyobb kihívást jelent, mint szöveges tartalom esetén. Nem elég a potenciálisan érdekes anyagok kiválasztása, hanem azok sorba rendezése is szükséges. A jó befogadási, hallgatási élmény alapvető egy ilyen tartalomszolgáltatás élvezhetőségéhez, elfogadottságához. A munka motivációját jelentő hangalapú kisközösségi média területén (1) beszédet és (2) zenét tartalmazó hanganyagokat is kezelni kell, ezek együtteséből kell a személyre szabott tartalmat kiválasztani.

A kihívás megoldásához a hang akusztikai és szöveges dimenzióját is érdemes vizsgálni a multimédiával kapcsolatos kutatásoknak megfelelően. Jelen munka az utóbbira vonatkozó első vizsgálatokat írja le, a beszédet tartalmazó hanganyagok leiratának, a dalszövegek és a hangzó beszédből való hangulatkinyerés vizsgálatával.

A munka felépítése a következő. A 2. fejezetben a tartalom személyre szabására vonatkozó szakirodalmat mutatjuk be. A 3. fejezetben a rádiós zenei szerkesztés gyakorlatának néhány kérdését vizsgáljuk. Ez alapján a nyelvi aspektusnál az érzelmek kezelése kulcsfontosságú, ezt a 4. fejezet tárgyalja. Az 5. fejezet a beszédfelismeréssel, a nyelvi reprezentáció szükségszerű minőségromlásával és ennek a hangulatfelismerésre való következményével foglalkozik, ezt követően összegezzük az eddigi eredményeket és tapasztalatokat, majd kitérünk a kutatás további terveire.

2 Kapcsolódó munkák

A tartalom személyre szabásának egyik alapeszköze az ún. ajánlórendszerek (recommender systems). Ezek célja olyan személyre szabott javaslatok készítése, amelyek az adott felhasználónak várhatóan tetszeni fognak. Ez történhet a tartalom- és az egyénprofil összehasonlításával, de akár – és sokszor ez a célszerűbb – kizárólag a többi felhasználó visszajelzései alapján az ún. collaborative filtering segítségével, pl. hasonló ízlésű felhasználók keresésével. Az ajánlórendszerek kutatási területén jelentős és gyakorlatreleváns eredményeket értek el, számos megközelítéssel [3]. Az ajánlórendszerek kimenete ugyanakkor nem rendezett úgy, hogy az lineáris médiához felhasználható legyen, erre a területen csak első próbálkozások ismertek [4,5].

A hanganyagok összeállítása folyamatos lejátszáshoz a szakirodalomban a lejátszási lista készítés (playlist generation) kifejezéshez kapcsolódik, egy aktuális áttekintés: [6]. A lejátszási listák készítésénél fegyelembe veendő tulajdonságokat többféleképpen csoportosítják, a szerzők meglátásaitól függően. Ezeket felhasználva [7] három szintet különböztet meg, és összegzi az egyes szinteken lényeges tulajdonságokat. Alulról felfelé haladva a szintek és tulajdonságok:

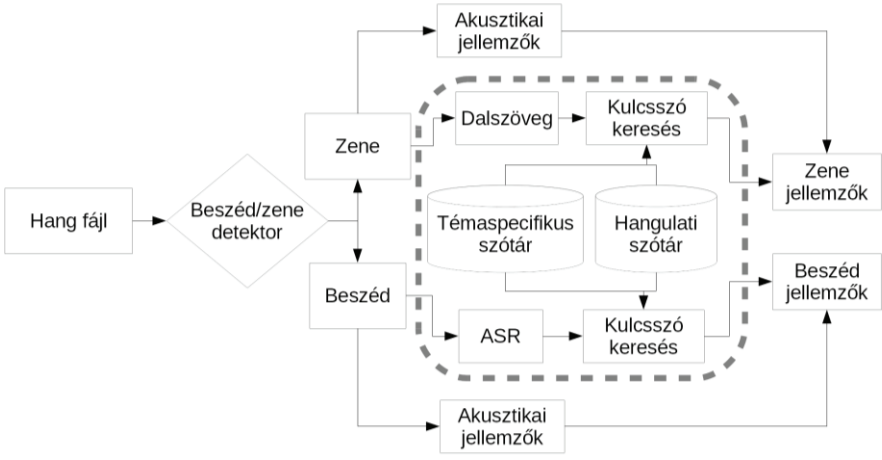
1. egyes dalok kiválasztása: frissesség-ismertség;
2. egymás utáni dalok kiválasztása: rendezettség-váratlanság;
3. a lejátszási lista egésze: koherencia-változatosság.

A fenti, lejátszási lista készítésére vonatkozó irodalom szinte kizárólag zenei lejátszási listákat kezel. A kevert beszéd-zene lejátszási listák készítése nyilvánvalóan különbözik ettől. Az utóbbiakra vonatkozó első kutatások [8] akusztikai szempontból vizsgálják az egymást követő beszéd-zene párokat. A [9]-ben leírt szabadalom sokkal átfogóbban, de a jelen munka szempontjából csak érintőlegesen foglalkozik a műsorválasztással, elsősorban források közötti átkapcsolással.

3 Megoldás tervezett felépítése

A rádiószerkesztés gyakorlatának vizsgálatához gyakorló zenei szerkesztőket kérdeztünk és a szöveges tartalom vizsgálatára egy, korábban a szakértői rendszerek felépítéséhez használt ún. korlátozott információs kísérletet (limited information experiment) [10] végeztünk. Ennek keretében előkészítettünk öt beszédet tartalmazó hanganyagot, majd elkészítettük ezek szöveges leiratát. A zenei szerkesztőnek elsőként csak a leiratot adtuk, és kértük, javasoljon zenét. Ezt követően az eredeti hanganyagot is lejátszottuk, azzal a kérdéssel, hogy mennyiben módosul a javasolt zene.

A tapasztalatok alapján csak a szöveg felhasználásával is jó minőségben tudott a szakember zenét ajánlani. A meghallgatás után ez csak kis mértékben módosult, pl. a beszéd tempója, vagy a beszélő kora, neme alapján. Az adatok kis száma ellenére is megerősödött így az a feltevés, hogy a beszédet tartalmazó hanganyagok szövegének jelentősége van a zenei szerkesztés szempontjából.



1. ábra. Hanganyagok jellemzőkinyerése automatikus beszéd-zene lejátszási lista készítéshez (ASR: beszédfelismerő).

A szakértőkkel való beszélgetésekből és a tesztekből kiderült, hogy szöveg-zene kapcsolódás esetén a legfontosabb illesztési szempont a hangulati jellemzők harmóniája, bizonyos speciális esetekben ezen kívül szerepet kap a témabeli illeszkedés. Ez többségében az ünnepkörök témáinál kívánatos. A cikkben leginkább a szöveg alapú

hangulatkinyeréssel foglalkozunk, a végső felhasználásban azonban a témabeli jellemzők is szerepet fognak játszani a feldolgozásban. Az 1. ábrán a hanganyagok jellemzőinek kinyeréséhez készített feldolgozási terv látható, kiemelve a cikk fókuszában található nyelvi feldolgozó részt.

4 Hangulatkinyerés írott szövegből

4.1 Dalszövegek és interjúátiratok előfeldolgozása

Bár a zenék hangulatának kinyeréséhez eredetileg kizárólag az akusztikai tulajdonságokat használták fel, a frissebb kutatások [11,12], illetve [13] kimutatták, hogy a dalszövegek felhasználása nagyban javítja a kategorizálás minőségét, sőt, sok esetben felülmúlja az akusztikai alapon végzett besorolás teljesítményét. Természetesen a dalszövegek felhasználását kívánatos ötvözni az akusztikai tulajdonságok feldolgozásával, a 3. fejezetnek megfelelően. Nem kizárólag azért, mert így érhető el a legjobb minőség, hanem azért is, mert nem minden zeneanyag szöveges, illetve nem minden anyag szövege hozzáférhető, nem megemlítve azt a tényt, hogy bizonyos esetekben a szöveg tartalmilag vagy terjedelmileg nem megfelelő a feldolgozásra.

Szintén a dalszövegek felhasználása mellett szól, hogy beszerzésük egyszerű és olcsó, szövegeik pedig többségében jól felhasználhatók a kívánt célra. Előbbi sajnos nem jellemző a beszédes tartalmakra. Bár megtalálhatók és hozzáférhetők interjúleiratok, ezek többségében nem elegendőek ahhoz, hogy az adatbázis legalább jó részét lefedő leiratokat szerezzünk, hiszen a tárolni kívánt beszédfelvételekhez aktualitásukból adódóan ritkán érhetőek el leiratok. Az ennek megoldására való kísérletek és a probléma részletezése az 5. fejezetben található. Ezen részben azokkal az esetekkel foglalkozunk, ahol – valamilyen szerencse folytán – rendelkezésre áll hivatalos leirat.

A dalszövegek feldolgozása bizonyos szempontból sajátos előkészítést igényelhet. Talán a legszembetűnőbb különbség az átlagos szövegek és a dalszövegek között (a sorok rövid volta mellett) az ismétlések jelzése. Ha olyan formában kívánjuk feldolgozni a szöveget, ahogyan az a dalban megjelenik, az ismétléseknek a szövegekben újra szerepelniük kell. Bár ezek a szövegek tulajdonképpen hivatalos leiratok és a különböző oldalakról kinyert anyagok egészen megegyezőek, az ismétlések jelzésére nincs kialakult módszer (‘repeat’, ‘4x’, ‘ref.’, ‘ref x2’), így nem is könnyű kezelni. Amellett, hogy többféle módon jelzik, meg kell állapítani az ismétlés hatókörét. Ez általában egy strófa vagy egy sor, attól függően, hol helyezkedik el a jelzés. Többségében ezek a jelzések az ismételni kívánt sor vagy sorok után helyezkednek el, azonban néha előttük. Előfordul, hogy a refrént sem írják ki, csak ennyivel jelzik: ‘ref.’ Ezek mind figyelmet érdemelnek, ha a feldolgozást valóban a teljes hangzó szövegen akarjuk végezni. Emellett azokat a részeket, amik nyilvánvalóan nem a dal szövegéhez tartoznak, mint az említett jelzések, sok esetben érdemes kiszűrni. Bár a dalszövegeket feldolgozó tanulmányok többségében nem foglalkoznak ezekkel a kérdésekkel, érdemes megfontolnunk, hogyan kezeljük az említett sajátosságokat. [14] és [15] normalizálja mindezeket és csak ezután következnek a feldolgozás további lépései.

Beszédleiratokban, legfőként interjúátiratok esetében meglehetősen sok kiszűrendő adat található, elég, ha csak arra gondolunk, hogy minden beszélőváltásnál kiírják az

éppen megszólaló nevét. Ezen kívül többnyire az átiratok elején szerepel egy kisebb bevezető, hogy mikor milyen műsorban hangzott el a beszélgetés egyéb kapcsolódó információk mellett.

Nagyobb anyag feldolgozásánál sok esetben szükséges, hogy felismerjük, milyen nyelvű a szöveg, legyen az akár dalszöveg, akár valamilyen beszédleirat. A kísérleteinkhez ezt a dalszövegeken el kellett végeznünk, mivel a használt adatbázisban különböző nyelven íródott dalok találhatók meg, a mérésekhez kizárólag az angol nyelvűeket használtuk fel. Az interjúátiratok esetében ezt nem tettük meg, mivel a BBC interjúátirataival dolgoztunk, melyek mind angol nyelvűek. Később szükséges lesz más nyelvű adatokkal is foglalkozni, így a nyelvfelismerés különösen fontos, leginkább az olyan további lépések esetében, ahol nyelvfüggő eszközt használunk (tipikusan ilyen a tokenizálás és lemmatizálás - amik szinte minden esetben szükségesek).

A dalszövegek felépítésének sajátosságához tartozik, hogy általában inkább csak mondattöredékeket tartalmaznak, nem pedig teljes mondatokat, írásjeleket ritkán találni, akkor is inkább figyelemfelhívó szerepben jelennek meg, mint mondathatárként. Emellett a mondatok, ha vannak is, sokszor sorokon átívelnek, így a mondatra bontás szinte lehetetlen, a Part of Speech taggelés igen nagy kihívás és kétséges, hogy megéri-e az eredmény az idő- és energiabefektetést.

4.2 A szövegekből való hangulatkinyerés módszerei

A hangulatkinyerés módszerét meglehetősen meghatározza, milyen típusú feldolgozást kívánunk végezni. Kategorizálni szeretnénk az anyagokat hangulati szempontból vagy egyszerűen metaadatként hozzáadjuk a hangulati jellemzőiket. Utóbbi többségében többdimenziós hangulati modell esetében jellemző. Ilyenkor a különböző dimenziókhoz tartozó értékeket rendelik az anyaghoz (ezek kétdimenziós modell esetén: pozitív-negatív érték, illetve az aktivitás mértéke). Jellemzőbb azonban a zenei hangulatkinyerésre, a hangulati kategóriába való besorolás. A használt kategóriarendszerek változatossága azonban igen nagy. A kategóriák számát illetően talán a legrobosztusabb a hat Ekman-féle [16] hangulati alapkategóriát használó, klasszikusnak mondható klasszifikálás. Ezek a következők: düh, undor, félelem, öröm, szomorúság, meglepettség. [15] 18 hangulati kategóriát különít el a last.fm címkéi alapján. Talán a legjellemzőbb továbbra is az Ekman-féle rendszer, valószínűleg részben azért, mert ez a pszichológiában klasszikusnak tekintett hangulati felosztás, illetve mert a jól felhasználható hangulati szótár, a WordNetAffect szintén ezt követi.

A dalszövegek hangulati feldolgozásában alapvetően kétféle módszer mérvadó leginkább. Az egyik a hangulati szótárakon alapuló, a másik pedig a gépi tanulásos módszer. Mindkettő esetében a meglévő és felhasználható eszközök igen függnek attól, milyen kategóriarendszert szeretnénk használni. Jónéhány hangulati szótár szabadon elérhető. Ilyenek a WordNetAffect (WNA) [17], a Linguistic Inquiry and Word Count (LIWC)[18], Affective Norms for English Words (ANEW) [19], illetve elérhető néhány más nyelvű szótár is, bár sok esetben ezeket az angol forrásokból fordítják.

A felügyelt gépi tanulás szintén jó eredményeket tud adni, hátránya azonban, hogy meglehetősen nagy méretű gold standard korpusz szükséges a tanuláshoz, amit igen ritkán adnak közre, ez a szükséges korpuszméret csökkenthető a hangulati szótárak felhasználásával. Felügyelt gépi tanulásnál többféle jellemzőt is figyelembe vesznek.

[14] elsősorban a TF*IDF módszert használja fel a tanuláshoz, de olyan globális jellemzőket is használnak, mint a szövegek és sorok hossza karakterben számolva.

A vizsgálatokban az Ekman-féle kategóriarendszert használtuk és a WordNetAffect hangulati szótárával dolgoztunk. A későbbiekben lehetséges, hogy az applikációban gyűjtünk mintát a hangulati kategorizálásra, így létrehozva egy gold standard korpuszt a későbbi módszerek kipróbálásához.

A hangulati szótárak használata bizonyítottan használható a dalszövegek hangulati kategorizálására, interjúátiratokra azonban eddig nem voltak kísérletek. Mivel a feldolgozás szempontjából a legegyszerűbb, ha az interjúszövegek és dalszövegek kategorizálása a lehető leghasonlóbban történik, arra voltunk kíváncsiak, hogy az interjúszövegek esetében is található-e annyi hangulati szó, hogy a találatok alapján bekategorizálhatók legyenek.

1. táblázat. A hangulati szavak gyakorisága dalszövegeknél (D: düh, U: undor, F: félelem, Ö: öröm, Sz: szomorúság, M: meglepettség).

id	D	U	F	Ö	Sz	M	összes hangulati szó (db)	lemmák száma (db)	hangulati szavak aránya a szövegben (%)
	hang. szavak aránya (%) a szövegben								
5710	0	0	0	8	0	92	13	194	6,70
…	…	…	…	…	…	…	…	…	…
5978	0	0	100	0	0	0	1	135	0,74
5980	0	0	0	36	27	36	11	368	2,99
						átl.:	**9**	**283**	**3,17**

Az 1. táblázat a dalszövegekben talált hangulati szavakat, a hangulati megoszlások arányában, illetve az összes hangulati szó számát a dalszövegben, a dalszöveg szószámát, illetve annak értékét, hogy a dalszöveg hány százalékát teszik ki hangulati szavak (utolsó oszlop). Látható, hogy ez az érték a vizsgált mintában 3,17%-os átlaggal szerepel. Ez azon okból fontos, hogy a hangulati kategorizálás általában kizárólag a talált hangulati szavak számából (illetve azok arányából) számítódik, azonban az, hogy milyen mértékben megbízható a történt besorolás, az függ attól, hogy milyen gyakorisággal szerepelnek hangulati szavak a szövegben. A kísérlet egy 200 elemű mintán folyt, a szövegeken nyelvfelismerést hajtottunk végre a Python langid.py [20] nyelvfelismerőjével és csak az angol szövegűeket dolgoztuk fel (174 szöveg). Ezeket a szövegeket a Python NLTK moduljának felhasználásával tokenizáltuk, majd lemmatizáltuk (WordNetLemmatizer). A szövegeket nem normalizáltuk, a refrének és ismétlések úgy szerepelnek, ahogy eredetileg voltak, ezt követően a szózsákokban kerestettük a hangulati szótárak elemeit. Az előfeldolgozásnál kipróbáltuk a stopszavak kiszűrését, de rontotta az eredményeket, minden valószínűség szerint a többszavas hangulati kifejezések esetében okozhatott ez hibát.

A 2. táblázat az előző táblázattal azonosan épül fel, a szövegfeldolgozás is ugyanazokat a lépéseket tartalmazta, azzal a különbséggel, hogy kiszűrtük a szövegből az olyan sorokat, melyek nem az interjú szövegéhez tartoztak (beszélők megjelölése,

interjú elejét és lezárását jelző címkék stb.). A BBC ‘Andrew Marr show’[1] műsoraiban lévő interjúk átiratait dolgoztuk fel, 282 átiratot elemeztünk a hangulati szavak megjelenésének szempontjából. Látható, hogy viszonylag sok hangulati szó szerepel az egyes átiratokban, ami a hangulati besorolás szempontjából nagyon jó, azonban látható, hogy a szövegek itt sokkal hosszabbak, mint a dalok esetében.

2. táblázat. A hangulati szavak gyakorisága interjúátiratoknál.

id	D	U	F	Ö	Sz	M	összes hangulati szó (db)	lemmák száma (db)	hangulati szavak aránya a szövegben (%)
	hang. szavak aránya (%) a szövegben								
1	0	0	8	29	14	49	49	1752	2,80
…	…	…	…	…	…	…	…	…	…
281	3	0	10	27	30	30	30	1997	1,50
282	0	0	0	36	27	36	30	1291	2,32
						átl.:	**28**	**2239**	**1,26**

Bár a hangulati kategorizálásnál nem veszik figyelembe a szövegek hosszát, csak a hangulati szavak megoszlásának arányát, a besorolás megbízhatósága függ attól, hogy az egész szövegben mekkora a hangulati szavak gyakorisága. A mérés itt igen nagy különbséget mutat a dalszövegek eredményeivel összehasonlítva, míg a dalszövegeknél a hangulati szavak százalékos aránya 3,17%, az interjúátiratoknál csak 1,26%. Ezek az értékek azt mutatják, hogy az interjúátiratoknál is használhatók a hangulati szótárak, azonban az eredmények kevésbé megbízhatók.

Érdekes még a feldolgozás szempontjából, hogy bár a kisebb gyakoriság miatt talán kevésbé megbízható az interjúátiratokra történő hangulati kategorizálás, nem volt olyan interjú, ahol ne találtunk volna hangulati szót. Tehát itt elvileg minden esetben lehetséges a hangulati szótár használata, a dalszövegek esetén azonban több alkalommal előfordult, hogy a szövegben egy hangulati szó sem szerepelt, így a csak hangulati szótár alapján történő besorolás ezen esetekben nem lehetséges. Lehetséges, hogy a hangulati szavak kis gyakorisága az interjúkban azok formális voltára vezethető vissza. Az interjúszövegek általában kifejtettebbek és sok olyan elemet tartalmaznak, amik kizárólag a társalgás fenntartására szolgálnak, pl.: szó átadása, köszönés stb., mely részek többnyire nem tartalmaznak hangulati szót.

5 Hangulatkinyerés hangzó szövegből

5.1 Bevezetés

A beszéd hangulatának tartalom alapján történő felismeréséhez a beszéd leiratára van szükség. Ez a legtöbb esetben nem áll rendelkezésre, ennek automatikus elkészítéséhez beszédfelismerő rendszert alkalmaznak. Az automatikus beszédfelismerő rendsze-

[1] http://www.bbc.co.uk/programmes/articles/3hshxFhHM4dKd3px6Q3NzRF/transcripts

rek a tanítóhalmaz és a kísérlethez használt felvétel paramétereitől függően (beszélők, akusztikai paraméterek, stb.) nagyon változatos eredményeket produkálhatnak a felismerési hibák szempontjából. A tanítóhalmazhoz hasonló felvételeken kevesebb hibát produkál, mint az attól jobban eltérőkön.

A hibák számszerűsítésére a szófelismerési hiba (word error rate) mértéket szokták használni. Minél kevesebb hibát okoz a beszédfelismerő, annál jobb eredményt kaphatunk a szótár alapú hangulatfelismerés szempontjából. Érdekes jelenség, hogy a hangulati szavak felismerésének növelésével nem lineárisan javul a hangulatfelismerés eredménye [21].

Kísérleteinkben a BBC 'In Touch' nevű műsorának fölvételeit és leiratait használtuk. Két módszert vizsgáltunk meg a hangulati szavak megtalálásához: a beszédfelismerő által felismert (legvalószínűbbnek számolt) szöveget használtuk a korábban ismertetett módon, illetve kulcsszókeresési eljárást alkalmaztunk a beszédfelismerő által generált, több valószínű útvonalat is tartalmazó hipotézis gráfon (lattice).

5.2 Felvételek

A felvételeket és a leiratokat a BBC In Touch[2] műsorának honlapjáról töltöttük le. Itt az utolsó öt adás mp3 formátumú hanganyaga és leirata érhető el. A hangfelvételek nem igényeltek különösebb előfeldolgozást, csak az elejéről kellett levágni a szerzői jogi információkat, valamint a beszédfelismerő tanulóhalmazához illeszkedve 16 KHz-es egycsatornás formátumba kellett konvertálni.

A leiratok általában pdf vagy rtf formában érhetőek el. Ezekben a szövegen kívül jelölve vannak a beszélőváltások is. A leiratozás nem gépi beszédfelismerést szem előtt tartva készült, ezért például a számok, a pénznemek nincsenek kiírva, az egyszerre beszélés, érthetetlen részek, stb. nincsenek jelölve, illetve ahol jelölve vannak ott sem egységes módon. Mivel nem volt kapacitás az ilyen jellegű hibák kijavítására, ezért minimális kézi előfeldolgozás után a problémák többségét (pl. számok átírása) automatikusan oldottuk meg.

5.3 Beszédfelismerés, kulcsszókeresés

A kísérletekhez a Kaldi beszédfelismerő rendszert használtuk [22]. A triphone modelleket változtatás nélkül a Kaldiban található előre elkészített TEDLIUM példakódokkal tanítottuk. Az akusztikai modell tanító adatbázisa a TEDLIUM[3] első verziója volt [23], nyelvi modellnek a CMUSphinx projekt által készített amerikai angol n-gram nyelvi modellt[4] használtuk.

A kulcsszókeresés és a beszédfelismerő által visszaadott legvalószínűbb útvonal esetén is ugyanazt a dekódolt hipotézis gráfot használtuk. Mindkét esetben tízszeres súlya volt a nyelvi modellnek az akusztikai modellhez képest. A beszédfelismerőnél a szófelismerési hibára (WER) 47,2%-ot kaptunk.

[2] http://www.bbc.co.uk/podcasts/series/intouch

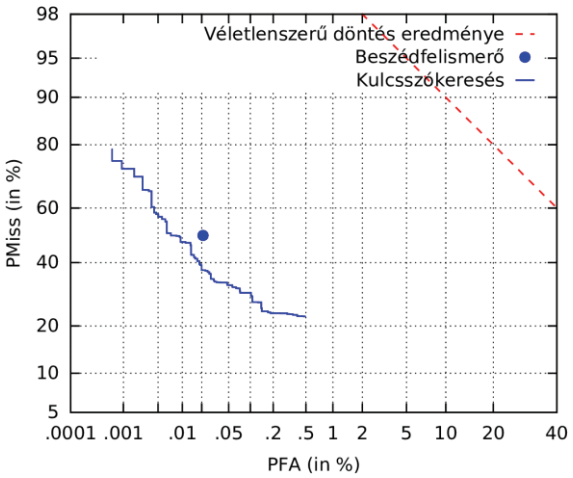
[3] http://www.openslr.org/resources/7/TEDLIUM_release1.tar.gz

[4] http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/US%20English%20Generic%20Language%20Model/cmusphinx-5.0-en-us.lm.gz/download

A kulcsszókeresésnél a Kaldiba beépített eljárást használtuk, ami a hipotézis gráfot indexeli a hivatkozott cikkben leírt módon [24]. Eredményül egy listát készít a megtalált kulcsszavakból, az elhangzás időpontjával és hosszával, valamint egy nulla és egy közötti értékkel, ami a találat bizonyosságát jelzi. Ez utóbbi alapján egy adott küszöbérték megadásával lehet eldönteni, hogy melyik találatokat tekintjük érvényesnek.

5.4 Eredmények

A két módszer – felismert szöveg és kulcsszókeresés – eredményét a nyers szövegen végzett kísérlet referenciaeredményével hasonlítottuk össze.



2. ábra. DET (Detection Error Tradeoff) görbe; a kulcsszókeresési módszer görbéje jó küszöbérték mellett kevesebb hibásan megtalált (PFA) és kevesebb nem megtalált (PMiss) kulcsszót eredményez (folytonos görbe), mint a beszédfelismerő által felismert szövegben való keresés (görbe feletti pont).

A módszerek eredményeit kulcsszókeresési algoritmusok kiértékeléséhez használt módszerek segítségével értékeltük ki [25], ezt az F4DE[5] eszközzel készített diagram (2. ábra) szemlélteti. Jó küszöbérték megválasztásával kevesebb hibásan megtalált (false alert) és kevesebb nem megtalált (miss) kulcsszót kaptunk a kulcsszókeresési módszerrel, mint a beszédfelismerő által felismert szöveg használatának esetén.

Mindkét módszernél kiszámoltuk a megtalált kulcsszavak érzelmi kategóriák szerinti eloszlását felvételenként, ezt hasonlítottuk a nyers szövegben történő keresésnél kapott eloszláshoz. Annak ellenére, hogy az előző kiértékelés alapján a kulcsszókeresést találtuk pontosabbnak, a hangulati szavak kategóriánkénti megoszlását vizsgálva a beszédfelismerő által felismert szövegben keresve az arányok jobban hasonlítottak a referenciaarányokhoz, mint bármelyik küszöbérték mellett a kulcsszókeresés esetén (3. táblázat).

[5] http://www.nist.gov/itl/iad/mig/tools.cfm

3. táblázat. a hangulati szavak eloszlása felvételenként a referencia, a beszédfelismerő által felismert szöveg és kulcsszókeresés esetén.

	Referencia (leirat)							Beszédfelismerő							Kulcsszókeresés						
Id	D	U	F	Ö	Sz	M	db	D	U	F	Ö	Sz	M	db	D	U	F	Ö	Sz	M	db
	%							%							%						
1021	3	0	14	63	6	14	35	4	0	16	60	9	11	45	0	0	22	61	9	9	23
1028	6	0	0	49	14	31	49	8	0	0	60	2	30	53	4	0	0	68	4	24	25
1104	19	0	0	44	6	31	48	12	0	0	47	17	25	60	19	0	0	51	11	19	37
1111	8	0	3	63	8	18	38	4	0	4	56	16	20	45	5	0	5	53	11	26	19
1118	4	0	7	52	9	28	46	0	0	8	46	12	33	48	0	0	7	52	11	30	27

A kulcsszókeresés rosszabb eredménye az eloszlás vizsgálatánál vagy a kevés adaton végzett vizsgálatnak köszönhető, vagy pedig annak, hogy a beszédfelismerés konzekvensen azonos arányú hibát okoz az egyes érzelmi kategóriák esetén. Annak eldöntéséhez, hogy melyik esetről van szó, nagyobb adathalmazon végzett további vizsgálatok szükségesek.

6 Összegzés és további tervek

A kutatás során kiderült, milyen módon képzelhető el az adatbázisban való hanganyagok összeillesztése és milyen szempontok játszanak szerepet a zene-beszéd listák létrehozásában. Kiderült, hogy a zene-beszéd illesztésnél a hangulati jellemzők a leginkább fontosak, megvizsgáltuk a hangulatkinyerés lehetséges módszereit és felmértük, mennyire lehet hatékony a hangulati szótár alapú kategorizálás interjúátiratok esetében. Ez alapján úgy tűnik, hogy bár interjúátiratok esetében kisebb a hangulati szavak gyakorisága (valószínűleg formális nyelvezetéből adódóan), mint az a dalszövegeknél szerepel, elég hangulati szót találni bennük a kategorizálásra, a dalszövegekkel ellentétben, ahol nem ritka, hogy átlagos hosszúságú szövegben egy hangulati szó sem szerepel, így ez alapján nem kategorizálható be.

Később azon általános esettel foglalkozunk, ahol a hangzó szövegből kell kinyernünk a hangulati szavakat, átirat hiányában. Az általunk vizsgált két módszer – a beszédfelismerő által felismert szövegben történő keresés és a kulcsszókeresés – jól közelíti a leiratban (referencia) való keresés eredményét. A kapott eredmények alapján nem lehet egyértelműen megállapítani, hogy melyik a jobb módszer. Ezt további, több adaton végzett vizsgálat segítségével lehetne kideríteni.

A továbbiakban figyelembe kívánjuk venni a szavak TF*IDF-értékét a hangulati szavaknál, így tehát azok a szavak nagyobb hangulati súllyal szerepelnének, melyek megkülönböztető szerepe nagyobb a szövegben, emellett részletes kidolgozásra kerül majd a hangulat akusztikai alapú felmérése, hiszen a végső felhasználásban a szöveges és akusztikai feldolgozás együtt szerepel majd.

Mivel szöveg-zene összekapcsolódásnál, bár csak bizonyos esetekben, de fontos a specifikus témák felismerése, ezt is kezelni kívánjuk. A megoldás a terv szerint hasonlóan működne, mint a hangulati szótárak alapján való besorolás, a különböző témákra, melyek fontosnak bizonyulnak (karácsony, újév stb.) egy-egy szótárat készítenénk, ezek alapján felismerve, hogy az anyag a témák valamelyikébe tartozik-e.

Hivatkozások

1. Kock, N.: The evolution of costly traits through selection and the importance of oral speech in e-collaboration. Electron. Mark. 19 (2009) 221–232
2. Lukacs, G., Pethesné, D. B., Madocsai, B.: Impact of Personalized Audio Social Media on Social Networks. In: XXXIII. Sunbelt Social Networks Conference of the International Network for Social Network Analysis Abstract Proceedings, Hamburg, Germany (2013) 210
3. Ricci, F., Shapira, B.: Recommender Systems Handbook. Springer (2011)
4. Zibriczky, D., Hidasi, B., Petres, Z., Tikk, D.: Personalized recommendation of linear content on interactive TV platforms: beating the cold start and noisy implicit user feedback. In: Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP), Montreal (2012)
5. Felfernig, A., Friedrich, G., Gula, B., Hitz, M., Kruggel, T., Leitner, G., Melcher, R., Riepan, D., Strauss, S., Teppan, E., Vitouch, O.: Persuasive Recommendation: Serial Position Effects in Knowledge-Based Recommender Systems. Persuasive Technology, Springer, Berlin / Heidelberg (2007) 283–294
6. Fields, B., Lamere, P.: Finding A Path Through The Jukebox – The Playlist Tutorial, ISMIR. Utrecht (2010)
7. Benyeda, I.: Zene–beszéd lejátszási listák készítésének nyelvtechnológiai vonatkozása. Pázmány Péter Katolikus Egyetem (2014)
8. Jani, M., Lukács, G., Takács, Gy.: Experimental Investigation of Transitions for Mixed Speech and Music Playlist Generation. In: Proceedings of ACM International Conference on Multimedia Retrieval, Glasgow, United Kingdom (2014) 392–398
9. Bull, W., Rottler, B.: Auto-station tuning, US Patent 8634944B2, Apple Inc, Cupertino, CA, US (2008)
10. Hoffmann, R.R.: The Problem of Extracting the Knowledge of Experts from the Perspective of Experimental Psychology 8 (1987) 53–67
11. Hu, X., Downie, J. S.: When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In: Proceedings of the 10th International Conference on Music Information Retrieval, Utrecht, The Netherlands (2010) 619–624
12. Hu, X., Chen, X., Yang, D.: Lyric-based Song Emotion Detection with Affective lexicon and Fuzzy Clustering Method. In: 10th International Society for Music Information Retrieval Conference (ISMIR 2009) (2009)
13. Mihalcea, R., Strapparava, C.: Lyrics, Music, and Emotions. In: Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn (2012) 590–599
14. van Zaanen, M., Kanters, P.: Automatic Mood Classification Using TF*IDF Based on Lyrics. In: Proceedings of the 11th International Society for Music Information Retrieval Conference (2010)
15. Hu, X., Downie, J. S., Ehmann, A. F.: Lyric Text Mining in Music Mood Classification. (2009) 411–416
16. Ekman, P.: Facial expression of emotion. Am. Psychol. 48 (1993) 384–392
17. Strapparava, C., Valitutti, A.: WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (2004) 1083–1086
18. Pennebaker, J.W., Martha, E.F., Booth, R.J.: Linguistic Inquiry and Word Count. Mahwah, NJ (2001)
19. Bradley, M. M., Lang, P. J.: Affective Norms for English Words (ANEW): Instruction manual and affective ratings, http://www.uvm.edu/~pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf (1999)

20. Lui, M., Baldwin, T.: langid.py: An Off-the-shelf Language Identification Tool. In: 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea (2012)
21. Chuang, Z.-J., Wu, C.-H.: Multi-Modal Emotion Recognition from Speech and Text. Comput. Linguist. Chin. Lang. Process. 9 (2004) 45–46
22. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding , Hilton Waikoloa Village, Big Island, Hawaii, US (2011)
23. Rousseau, A., Deléglise, P., Estève, Y.: TED-LIUM: an Automatic Speech Recognition dedicated corpus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey (2012)
24. Can, D., Saraclar, M.: Lattice Indexing for Spoken Term Detection. IEEE Trans. Audio Speech Lang. Process. 19 (2011) 2338–2347
25. Fiscus, J. G., Ajot, J., Garofolo, J. S., Doddingtion, G.: Results of the 2006 spoken term detection evaluation. In: Proc. SIGIR (2007) 51–57